Network Harms

Andy Z. Wang†

When it comes to data, the whole is greater than the sum of its parts. There may be millions of people with the same birthday. But how many also have a dog, drive a red car, and have two kids? The more that data is aggregated, the more identifying, and thus sensitive, it becomes. In recognition of this principle, the law has developed safe harbors for firms that take steps to prevent aggregation of the data they sell. A firm might, for instance, anonymize its data by removing identifying information. But as the science academy has shown, a wide array of de-anonymization techniques largely renders such efforts moot, allowing motivated actors to link data back to its data subject even after data is anonymized.

Today, data collection is ubiquitous. Data brokers—firms that collect, process, and sell the data of individuals with whom they have no direct business relationship—are a major player in this ecosystem. Courts have traditionally conceived of the harms that arise from data brokering as discrete harms divorced from other activity occurring in the larger data ecosystem. But that judicial conception overlooks a crucial intuition: the magnitude of harm arising from one broker's activities depends on what data other brokers in the network are selling. De-anonymization techniques often depend on cross-referencing external data to make internal inferences. Furthermore, a motivated actor can purchase multiple datasets from multiple brokers, employ de-anonymization techniques to overcome barriers to aggregation, and aggregate as they please. The consolidated dataset would represent a far larger privacy harm than if those datasets were to be owned in isolation. These "network harms" have thus far been underexplored.

In the absence of meaningful legislation, this Comment urges a turn to the courts. Section 5 of the Federal Trade Commission Act empowers the Federal Trade Commission (FTC) to prevent "unfair or deceptive acts or practices." In recent years, the FTC has begun to employ these statutory powers to reach the activities of several data brokers. This Comment offers a framework for courts to incorporate network harms in § 5 suits. Doing so would provide a more accurate descriptive account of brokering harms and also extend the reach of the FTC's theories to data brokers whose activities are not grossly negligent but nevertheless harmful.

---

† B.S. 2022, San Jose State University; J.D. Candidate 2025, The University of Chicago Law School. I would like to thank Professor Omri Ben-Shahar for his tremendous guidance and advice. Thank you to the editors and staff of the *University of Chicago Law Review* for their tireless editing support. A special thank you to Eric Haupt, Jack Brake, Karan Lala, Tanvi Antoo, Luke White, Jake Holland, Bethany Ao, Emilia Porubcin, Benjamin Wang, and Anastasia Shabalov for their invaluable insights and contributions along the way.

INTRODUCTION

In July 2021, a top Catholic Church official resigned after a Catholic news site outed him by purchasing his cell-phone data from a data broker, exposing his visits to gay bars and use of Grindr.[1] The incident—one of many—"highlighted the dangers of the large, shadowy, and unregulated data brokerage industry selling Americans' real-time locations to the highest bidder."[2] Data brokers specialize in buying, aggregating, and selling the

---

[1] *See* Justin Sherman, *Data Brokers Know Where You Are—and Want to Sell That Intel*, WIRED (Aug. 23, 2021) [hereinafter Sherman, *Brokers Know Where You Are*], https://www.wired.com/story/opinion-data-brokers-know-where-you-are-and-want-to-sell-that-intel/; Michelle Boorstein, Marisa Iati & Annys Shin, *Top U.S. Catholic Church Official Resigns After Cellphone Data Used to Track Him on Grindr and to Gay Bars*, WASH. POST (July 21, 2021), https://perma.cc/V45S-RKYU.

[2] Sherman, *Brokers Know Where You Are*, *supra* note 1.

personal data of individuals with whom they have no direct relationship. This data is valuable[3] to corporate interests because wielding it enables them to "target consumers with highly personalized offers, recommendations[,] and information."[4] But as the former Catholic priest's story illustrates, that data can also be used for more sinister purposes.[5] In fact, the current state of affairs practically invites it.

Today, data brokers are largely free to sell highly sensitive data to the highest bidder. Even when identifying information, like someone's name, is omitted, most of this data can be linked back to its original source (i.e., data subject) through clever inferences.[6] To make matters worse, individuals often do not know that their data is being collected, let alone that it is being sold.[7]

In the United States, there is no comprehensive federal law governing data brokers.[8] In fact, the United States lacks a comprehensive federal law governing data privacy altogether.[9] Rather, data privacy at the federal level is governed by a patchwork of decades-old laws that protect narrow categories of sensitive data in sector-specific circumstances. For instance, the Health Insurance Portability and Accountability Act[10] (HIPAA) regulates how healthcare providers can use a patient's personal health

---

3 The global data broker market was valued at $319 billion in 2021 and is projected to reach $545 billion in 2028. *See Global Data Broker Market Size, Share, Opportunities, COVID-19 Impact, and Trends by Data Type (Consumer Data, Business Data), by End-User (BFSI, Retail, Automotive, Construction, Others), and by Geography—Forecasts from 2023 to 2028*, KNOWLEDGE SOURCING INTELLIGENCE (June 2023), https://perma.cc/3JQ7-RG8L.

4 Brian Fung, *DOJ Will Hire More Data Experts to Scrutinize Digital Monopolies, Antitrust Chief Says*, CNN (Mar. 6, 2023), https://perma.cc/8ZA5-QJXA.

5 For a far less sympathetic instance of identity exposure via data broker, see Dhruv Mehrotra & Dell Cameron, *Jeffrey Epstein's Island Visitors Exposed by Data Broker*, WIRED (Mar. 28, 2024), https://www.wired.com/story/jeffrey-epstein-island-visitors-data-broker-leak/.

6 *See infra* Part II.B.

7 *See* Justin Sherman, *Data Brokers and Sensitive Data on U.S. Individuals*, DUKE SANFORD CYBER POL'Y PROGRAM 2 (2021) [hereinafter Sherman, *Sensitive Data*], https://perma.cc/W9RV-VR6G ("Consumers do not necessarily know that the data about them is being collected.").

8 *Id.* ("Data brokerage . . . is a virtually unregulated practice in the United States.").

9 *See* Conor Murray, *U.S. Data Privacy Protection Laws: A Comprehensive Guide*, FORBES (Apr. 25, 2023), https://www.forbes.com/sites/conormurray/2023/04/21/us-data-privacy-protection-laws-a-comprehensive-guide ("Data privacy in the United States is notably different than in the [European Union], which has a comprehensive data privacy law.").

10 Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered sections of the U.S. Code).

data,[11] and the Gramm-Leach-Bliley Act[12] (GLBA) requires financial institutions to safeguard sensitive data and disclose how they use customer data.[13] But while these statutes offer forceful protections so far as they extend, many entities and categories of sensitive data fall beyond their reach.

Location data is one example—the absence of comprehensive federal protections has invited a parade of abuses. Government agencies like U.S. Immigration and Customs Enforcement (ICE), for instance, have used data brokers to track suspected illegal immigrants, even when sanctuary laws empower a municipality to refuse the agency's requests for information.[14] Some data brokers even sell the location data of people who visit abortion clinics, which, in the wake of *Dobbs v. Jackson Women's Health Organization*,[15] may be weaponized to prosecute women who seek abortions.[16] Furthermore, because these sectoral privacy laws were passed in the 1980s and 1990s, they "fail to cover the relatively new phenomenon of online data brokers . . . that have only materialized in the last [twenty] years."[17]

The absence of a comprehensive federal law has spawned a patchwork of state-led efforts to protect data privacy. In 2018, following the Cambridge Analytica scandal,[18] California enacted the California Consumer Privacy Act[19] (CCPA), described at the time of its enactment as the "broadest, most overarching privacy law

---

11 *See generally id.*

12 Pub. L. No. 106-102, 113 Stat. 1338 (1999) (codified as amended in scattered sections of the U.S. Code).

13 *See generally id.*

14 *See* Saja Hindi & Elizabeth Hernandez, *ICE Uses Private Data Companies to Circumvent Colorado "Sanctuary" Laws, New Report Says*, DENVER POST (Apr. 22, 2022), https://www.denverpost.com/2022/04/21/ice-private-data-colorado-sanctuary-laws-report/; Cristiano Lima-Strong & Aaron Schaffer, *ICE's Use of Data Brokers to 'Go Around' Sanctuary Laws Under Fire*, WASH. POST (July 27, 2022), https://perma.cc/JB9A-YHW5.

15 142 S. Ct. 2228 (2022).

16 Joseph Cox, *Data Broker Is Selling Location Data of People Who Visit Abortion Clinics*, VICE (May 3, 2022), https://perma.cc/E9GU-DDX6.

17 Caitriona Fitzgerald, Kara Williams & R.J. Cross, *The State of Privacy: How State "Privacy" Laws Fail to Protect Privacy and What They Can Do Better*, EPIC 13 (Feb. 2024), https://perma.cc/A394-LTFB.

18 Sara Morrison, *California's New Privacy Law, Explained*, VOX (Dec. 30, 2019), https://perma.cc/P2CW-QEJJ (explaining how Cambridge Analytica, a political consulting firm, exploited Facebook's developer tools to access and collect data from eighty-seven million profiles largely without notice).

19 2018 Cal. Stat. 1807 (codified as amended at CAL. CIV. CODE §§ 1798.100–.199.100).

passed in the U.S."[20] The Act and its progeny[21] grant California consumers mechanisms to assert control over how their personal data is collected, used, and sold.[22] Following in California's footsteps, thirteen other states have enacted their own comprehensive data privacy laws.[23] Four states—Vermont, Texas, California, and Oregon—have enacted laws specifically targeting data brokers.[24] Unfortunately, however, the resulting patchwork regime leaves much to be desired. Not only have individual states adopted imperfect laws,[25] but data's amorphousness renders it impractical to police data transfers across state lines.

At bottom, Congress has fallen short. Industry lobbying has halted comprehensive federal legislation in its tracks, the vast majority of state data privacy laws are written by the very industry giants those laws seek to regulate,[26] and state data broker laws remain relatively scarce. In light of this state of affairs, this Comment offers two suggestions. First, courts should adopt a more pragmatic conception of harm in the data broker context. The traditional judicial approach largely treats data brokers as discrete sellers selling to discrete buyers, and the act of sale as producing discrete harms. But the data brokerage ecosystem is interconnected—different brokers can collect overlapping data, and repeat buyers can purchase and aggregate data from multiple

---

[20] Stuart L. Pardau, *The California Consumer Privacy Act: Towards a European-Style Privacy Regime in the United States?*, 23 J. TECH. L. & POL'Y 68, 73 (2018).

[21] *See* CAL. CIV. CODE §§ 1798.100–.199.100; 2023 Cal. Stat. 6632 (codified in scattered sections of CAL. CIV. CODE).

[22] Gopal Ratnam, *Push for Federal Data Privacy Law Grows as Rights Vary by State*, ROLL CALL (Jan. 17, 2024), https://perma.cc/8JRM-VC35. The Act also confers a private right of action for consumers to directly sue companies over data breaches involving personal information. *Id.*

[23] The states that have passed comprehensive data privacy laws are California, Colorado, Connecticut, Delaware, Indiana, Iowa, Kentucky, Montana, New Hampshire, New Jersey, Oregon, Tennessee, Utah, and Virginia. *See* Andrew Folks, *US State Privacy Legislation Tracker*, INT'L ASS'N PRIV. PROS. (Apr. 12, 2024), https://perma.cc/LZ78-BHXQ.

[24] *See* VT. STAT. ANN. §§ 2430–31, 2447, 2466 (2023); TEX. BUS. & COM. CODE ANN. § 509.001 (West 2023); CAL. CIV. CODE § 1798.99.80 (West 2024); OR. REV. STAT. ANN. § 646A.593 (West 2024).

[25] *See, e.g.*, *California Consumer Privacy Act (CCPA)*, OFF. ATT'Y GEN. (Mar. 13, 2024), https://perma.cc/VT8M-HRSZ ("The CCPA's definition of 'personal information' does not include information lawfully made available from government records, which are often sources used by data brokers.").

[26] *See* Alfred Ng, *Privacy Bill Triggers Lobbying Surge by Data Brokers*, POLITICO (Aug. 28, 2022), https://perma.cc/8EXD-5C9K. A nationwide approach to data privacy has seen success overseas. *See generally* Commission Regulation 2016/679, art. 4, 2016 O.J. (L 119). The absence of a similar law in the United States is largely attributable to lobbying. *See* Alfred Ng & Maddy Varner, *The Little-Known Data Broker Industry Is Spending Big Bucks Lobbying Congress*, THE MARKUP (Apr. 1, 2021), https://perma.cc/67MD-EJ26.

brokers. The current framing fails to account for these "network harms,"—the harms that arise when a buyer independently aggregates data after purchasing from multiple brokers. A network-sensitive conception of privacy harm would recognize that the harm arising from one broker's sales depends at least partly on what data other brokers in the network are offering for sale.

Attempts to prevent such aggregation also fall short. This is because data is synergistic. That is, when it comes to data, the whole is greater than the sum of its parts.[27] Even when identifying information is omitted from a dataset (i.e., anonymized), motivated actors can employ sophisticated *de*-anonymization techniques, cross-referencing datasets to identify individuals even when such identification would have been otherwise impossible when looking at those datasets separately.[28] Taken together, a motivated actor can purchase multiple datasets from different brokers, capitalize on data's synergistic nature via de-anonymization techniques to overcome barriers preventing aggregation, then aggregate as they please. The law has not responded to these risks, and the ensuing absence of liability underincentivizes responsible data brokering.[29] Adopting this broader conception of harm in the data brokerage context would realign the incentives for responsible brokering and recognize that data becomes exponentially more powerful—and therefore potentially harmful—the more of it there is.[30]

---

[27] This principle is known in the Fourth Amendment context as the "mosaic theory." *See* Matthew B. Kugler & Lior Jacob Strahilevitz, *Actual Expectations of Privacy, Fourth Amendment Doctrine, and the Mosaic Theory*, 2015 SUP. CT. REV. 205, 205 ("The mosaic theory of the Fourth Amendment holds that, when it comes to people's reasonable expectations of privacy, the whole is greater than the sum of its parts."); *see also* U.S. Dep't of Just. v. Reps. Comm. for Freedom of the Press, 489 U.S. 749, 764–65 (1989) (holding that although the "individual events" in data subjects' criminal records were "matters of public record," those data subjects had a privacy interest in the aggregated "whole" distinct from their interest in the "bits of information" considered individually); United States v. Maynard, 615 F.3d 544, 558 (D.C. Cir. 2010) ("[T]he whole of one's movements is not exposed *constructively* even though each individual movement is exposed, because that whole reveals more—sometimes a great deal more—than does the sum of its parts.").

[28] *See De-anonymization*, TECHTARGET (May 2015), https://perma.cc/5UQ6-43A9; *infra* Part II.B.

[29] *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1740–41 (2010).

[30] *See* Aaron Fluitt, Aloni Cohen, Micah Altman, Kobbi Nissim, Salome Viljoen & Alexandra Wood, *Data Protection's Composition Problem*, 5 EUR. DATA PROT. L. REV. 285, 287 (2019) (recognizing this characteristic as "[c]omposition effects[:] the cumulative results of multiple uses of data vis-a-vis data privacy"); *see also id.* at 292 (describing these cumulative effects as a "tyranny of small decisions: although each step seems small, together they bring society over a cliff").

Second, existing laws serve as a suitable vehicle to incorporate such network harms. Specifically, this Comment urges a turn to courts and the Federal Trade Commission (FTC). Section 5 of the Federal Trade Commission Act[31] (FTCA) empowers the FTC to prevent "unfair or deceptive acts or practices in or affecting commerce."[32] The FTC has applied this power to reach everything from telemarketing schemes[33] to deceptive shaving cream commercials.[34] Recently, the FTC has been testing this theory against data brokers in federal court and administrative tribunals with varying success.[35] Section 5 not only provides a path forward in resisting the commodification of sensitive data, but also provides a vehicle for applying a broader, network-sensitive conception of privacy harm. This Comment offers a framework for courts to conceive of these network harms and posits ideas on how to enhance these theories by emphasizing the risks of de-anonymization.

This Comment proceeds as follows. Part I introduces the underregulated data brokerage industry and highlights the dangers it poses. Part II explores inherent characteristics of data that make it difficult for regulators to accurately conceive of data brokering harms. It further demonstrates how, in light of these characteristics of large datasets, the act of aggregating data generates the most severe privacy harms. In conceiving of brokering harms, regulators should consider the network of data brokers that enable these sorts of aggregations. Part III explores potential solutions in doctrine concerning the FTC's authority to prevent "unfair acts and practices" and an ongoing case against Kochava, a data broker, that marks the first time the FTC has tested this theory in federal court. Part IV offers a framework that incorporates these network risks into the judicial conception of harm before applying the framework to Kochava.

I. THE STATUS QUO: FEDERAL LAW LACKS COMPREHENSIVE DATA PRIVACY PROTECTIONS

"Data brokerage—broadly, the practice of buying, aggregating, [and] selling" the data of individuals with whom the broker

---

31 15 U.S.C. §§ 41–58.

32 *Id.* § 45(a)(1).

33 *See, e.g.*, FTC v. Windward Mktg., Inc., 1997 WL 33642380, at *1 (N.D. Ga. Sept. 30, 1997).

34 *See, e.g.*, FTC v. Colgate-Palmolive Co., 380 U.S. 374, 376 (1965).

35 *See Privacy and Security Enforcement*, FED. TRADE COMM'N, https://perma.cc/MU8B-9R4G.

has no direct relationship—"is a virtually unregulated practice in the United States."[36] Collectively, the data brokerage industry collects and sells data on "virtually every American."[37] Data brokers deal in data about individuals' demographics, political beliefs, addresses, geolocations, health conditions, financial well-being, and "lifestyle characteristics (such as travel, media consumption, and mobile app usage)."[38] A single data broker "might have anything from a few data points . . . to hundreds or thousands of data points about a single person."[39] How did this situation arise, and what are its implications? First, this Part provides an overview of the data brokerage industry—what data brokers do and how they do it. Second, it articulates the status quo—systematic underregulation—and the dangers that arise when data brokers can sell whatever they want to whomever they want.

A. How Do Data Brokers Amass Data, and What Do They Do With It?

Data brokers collect data in three main ways: directly, indirectly, and by inference.[40] Brokers collect data *directly* by, for example, entering contracts with app developers to include the broker's data-siphoning software directly in their apps.[41] They also collect data *indirectly* by "scraping public records . . . , gathering data from real-time bidding networks for online ads," and buying data from first-party collectors.[42] Perhaps most interestingly, brokers collect

---

[36] Sherman, *Sensitive Data*, *supra* note 7, at 2. The term has been defined thus far by four state data broker laws in California, Oregon, Texas, and Vermont. *See supra* note 24. California's law, for instance, defines "data broker" as a "business that knowingly collects and sells to third parties the personal information of a consumer with whom the business does not have a direct relationship." *See* CAL. CIV. CODE § 1798.99.80 (West 2024).

[37] Justin Sherman, David Hoffman, Spencer Reeves, Aden Klein, Brady Allen Kruse, Alistair Simmons & Hayley Barton, *Response from Duke University's Data Brokerage Research Project: Consumer Financial Protection Bureau (CFPB) Request for Information Regarding Data Brokers and Other Business Practices Involving the Collection and Sale of Consumer Information*, DUKE SANFORD CYBER POL'Y PROGRAM 2 (July 2023), https://perma.cc/9SMJ-YPVP.

[38] *Id.*

[39] *Id.*

[40] *See id.* at 5.

[41] *See id.*

[42] Sherman et al., *supra* note 37, at 5. *See* Jon Keegan & Alfred Ng, *Gay/Bi Dating App, Muslim Prayer Apps Sold Data on People's Location to a Controversial Data Broker*, THE MARKUP (Jan. 27, 2022), https://perma.cc/MZ5T-5UUZ (compiling apps that sold data to data brokers).

data *by inference* by making predictions about individuals' characteristics based on the existing data they own.[43] For instance, if a data broker knows that an individual collects vinyl records and frequently posts about their collection to social media, they might infer that the individual owns a record player.[44] This is an easy inference that barely requires pen and paper, let alone a computer, to make. But today, aided by improving algorithmic and computing capabilities, data brokers can make complex inferences faster than ever.[45] A data broker might "follow[ ] individuals' smartphone geolocation patterns over time to learn about their visits to home, work, retail stores, medical facilities, gay bars, and places of worship."[46] And as discussed in Part II.B, these techniques can also be applied to de-anonymize data that is otherwise anonymous. Because data naturally accumulates over time, and due to recent technological advancements in machine learning, it is likely that techniques for "collecting" data by inference are becoming more efficient, effective, and lucrative.[47]

Data brokers then package and sell this data to a wide range of different clients, from insurance companies to political campaigns.[48] There is also cross-pollination in the data brokerage industry—an FTC report from 2014 found that "[s]everal . . . data brokers share the same sources" and that the majority of the studied brokers "buy from or sell information to each other."[49]

B. The Dangers of an Underregulated Data Brokerage Industry

In many circumstances, consumers benefit when third parties own some of their personal data. Data on someone's spending behaviors, for instance, enables banks to more effectively detect

---

43 Sherman et al., *supra* note 37, at 5.

44 *But see* Jaime Marconette, *Top Entertainment Trends for 2023: What the Data Says*, LUMINATE 9 (Mar. 22, 2023), https://perma.cc/3HD4-H5PA ("50% of consumers who have bought vinyl in the past twelve months [do not] own a record player.").

45 *See* Sherman et al., *supra* note 37, at 5.

46 *Id.* at 4.

47 *Cf.* Yash Sherry & Neil C. Thompson, *How Fast Do Algorithms Improve?*, 109 PROCEEDINGS IEEE 1768, 1769 (2021) (finding that, "for moderate-sized problems, 30%–43% of algorithmic families had improvements comparable or greater than those that users experienced from . . . hardware advances").

48 *See* FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY 39–40 (2014) [hereinafter FTC, DATA BROKERS] (charting data broker clients by product type and industry sector).

49 *Id.* at 14.

fraudulent charges.[50] Data on how commuters travel via car, bike, bus, or rail "provide[s] answers on [the] best infrastructure types to invest in," resulting in better cities.[51] Data on shopping behaviors can be used to tailor advertising, which consumers have come to expect from platforms.[52] To the extent that the data brokerage industry enables and facilitates these socially beneficial uses of data, these brokers play a valuable role in the data ecosystem.[53]

The question, then, is not whether data brokers should be banned altogether, but rather how society can maximize data brokers' welfare-enhancing role in the ecosystem while minimizing the risks of harm. "Major data brokerage firms are presently offering reams of data on U.S. individuals for sale . . . ."[54] This data often includes highly sensitive data, such as "race, ethnicity, gender, sexual orientation, immigration status, income level, and political

---

[50] *See Big Data Analytics: A Fraud Prevention Game Changer*, FRAUD.NET, https://perma.cc/TB6S-99XS.

[51] Pranab K. Roy Chowdhury, Susanna H. Sutherland, Kathleen M. Ernst, Alexander Pawlowski, Erik H. Schmidt, Janna R. Caspersen, Ziliang Zhao & Budhendra L. Bhaduri, *Big Data in Emerging Cities*, *in* BIG DATA FOR REG'L SCI. 271, 283 (Laurie A. Schintler & Zhenhua Chen eds., 2017).

[52] *See* Nidhi Arora, Daniel Ensslen, Lars Fiedler, Wei Wei Liu, Kelsey Robinson, Eli Stein & Gustavo Schüler, *The Value of Getting Personalization Right—or Wrong—Is Multiplying*, MCKINSEY & CO. (Nov. 12, 2021), https://perma.cc/7A6N-34XC ("Seventy-one percent of consumers expect companies to deliver personalized interactions."); *see also* Kristen O'Shaughnessy, D. Daniel Sokol, Jaclyn Phillips & Nathan Swire, *Big Data, Little Chance of Success: Why Precedent Does Not Support Anti-Data Theories of Harm*, CPI ANTITRUST CHRON. 2 (July 2022) https://perma.cc/ZY35-PZAF (observing that "Big Data . . . has enabled extraordinary innovation, creating a number of benefits, including free products and greater efficiencies"); Omri Ben-Shahar, *Privacy Protection, At What Cost? Exploring the Regulatory Resistance to Data Technology in Auto Insurance*, 15 J. LEGAL ANALYSIS 129, 136 (2023) (evaluating the negative effects of resisting "usage-based [car] insurance" that relies on personal data, which induces safer driving, reduces fatal accidents, and results in more affordable and fair premiums).

[53] Furthermore, data brokers potentially democratize the data trade by enabling smaller players who do not have the same first-party collecting capabilities as larger companies like Google to compete in the market. For example, companies "need access to personal data" to compete and innovate. MAURICE E. STUCKE, BREAKING AWAY: HOW TO REGAIN CONTROL OVER OUR DATA, PRIVACY, AND AUTONOMY 165–66 (2022). For an example of data brokers democratizing data access for nonprofit purposes, consider SafeGraph, a data broker that enables academics to access its data. *See, e.g.*, *SafeGraph Partners with Dewey to Democratize Access to Data for Academics*, SAFEGRAPH (Sept. 2, 2022), https://perma.cc/89ZQ-Z3YP. Basic economic theory purports that this increased competition enhances consumer welfare. *See, e.g.*, Heather Boushey & Helen Knusdsen, *The Importance of Competition for the American Economy*, THE WHITE HOUSE (July 9, 2021), https://www.safegraph.com/blog/safegraph-partners-with-dewey ("Basic economic theory demonstrates that when firms have to compete for customers, it leads to lower prices, higher quality goods and services, greater variety, and more innovation.").

[54] *See* Sherman, *Sensitive Data*, *supra* note 7, at 2.2.

preferences . . . ."[55] If left unregulated, brokers may not exercise due care when transacting with this sensitive data.

Three observations highlight the problems with this status quo. First, data brokers are buyer agnostic—they can sell sensitive data to virtually whomever they would like. For instance, brokers that deal in location data—often captured innocuously through mobile apps—frequently sell this data to third parties.[56] If this data falls into the wrong hands, it can be used to stalk and harass.[57] Moreover, data on race, gender, sexual orientation, and immigration status can enable discriminatory policing and surveillance, presenting a particularly heightened danger for socially marginalized groups.[58] Some have even argued that data brokers threaten national security.[59] "[V]irtually nothing in current U.S. law limits [ ] selling [ ] data [on U.S. individuals] to a range of actors, from insurance firms to U.S. law enforcement agencies to foreign entities."[60] "[T]here is little transparency" into these transactions.[61] One report found that some large data brokers "explicitly advertise data on current and/or former U.S. military personnel."[62] And, surprisingly, "virtually nothing in U.S. law prevents data brokers from selling information on U.S. individuals to foreign entities."[63] The data these brokers offer for sale—"spanning everything from financial transaction histories and internet browsing patterns to travel interests and support for political causes and organizations—could be used by foreign entities for a range of national security-damaging activities" such as foreign surveillance or scams.[64] Foreign governments can also use this data to "micro-target[ ] individuals with election disinformation," like the Russian Internet Research Agency did to Black communities during the 2016 U.S. presidential election.[65]

---

55 *Id.* at 9.

56 *See, e.g.*, *FTC Sues Kochava for Selling Data that Tracks People at Reproductive Health Clinics, Places of Worship, and Other Sensitive Locations*, FED. TRADE COMM'N (Aug. 29, 2022) [hereinafter *FTC Sues Kochava*], https://perma.cc/NZ2J-LATE.

57 *Id.*

58 *See, e.g.*, Sherman, *Sensitive Data*, *supra* note 7, at 10.

59 *See id.* at 10–11.

60 *Id.* at 2.

61 *Id.* at 11.

62 *Id.* at 10.

63 Sherman, *Sensitive Data*, *supra* note 7, at 11; *see also Fact Sheet: President Biden Issues Executive Order to Protect Americans' Sensitive Personal Data*, THE WHITE HOUSE (Feb. 28, 2024), https://perma.cc/E33P-A7H5.

64 Sherman, *Sensitive Data*, *supra* note 7, at 11.

65 *Id.*

And it is not just foreign governments. U.S. government agencies like ICE and the FBI can purchase data from data brokers "without warrants, public disclosures, or robust oversight."[66] ICE, for one, has accessed "a private database containing hundreds of millions of phone, water, electricity and other utility records while pursuing immigration violations."[67] The agency also searched LexisNexis's massive database of personal information over 1.2 million times in a seven-month period in 2021.[68] As discussed in the Introduction, ICE has also used data brokers to circumvent sanctuary laws,[69] and some data brokers are selling the location data of people who visit abortion clinics, potentially exposing them to harassment and prosecution.[70]

Government agencies are also not the only entities that can misuse this data. A recent report found that "there are data brokers which advertise and are willing and able to sell data concerning Americans' highly sensitive mental health information" in ways that do not violate HIPAA.[71] "[C]ompanies outside the narrow scope of HIPAA, from data brokers to period tracking apps, can legally sell Americans' health-related information, and they do, from a list of your surgical procedures to your mental health conditions."[72]

Third, because data brokers often do not collect the data themselves, they often have little incentive to care about who they sell their data to and what that buyer will do with the data. For example, LexisNexis "advertises a capability to search an individual and identify whether they are active-duty military," which can expose these individuals to espionage attempts, scams, or

---

66 Sherman, *Brokers Know Where You Are*, *supra* note 1; *see also* Matthew Tokson, *Government Purchases of Private Data*, 59 WAKE FOREST L. REV. 269, 283–88 (2024) (collecting instances where federal agencies and police departments purchased data from brokers and other vendors).

67 Drew Harwell, *ICE Investigators Used a Private Utility Database Covering Millions to Pursue Immigration Violations*, WASH. POST (Feb. 26, 2021), https://perma.cc/9C4R-JREA.

68 Sam Biddle, *ICE Searched LexisNexis Database Over 1 Million Times in Just Seven Months*, THE INTERCEPT (June 9, 2022), https://perma.cc/E46C-DEA8.

69 Lima-Strong & Schaffer, *supra*, note 14.

70 Cox, *supra* note 16.

71 Joanne Kim, *Data Brokers and the Sale of Americans' Mental Health Data*, DUKE SANFORD CYBER POL'Y PROGRAM 2 (Feb. 2023), https://perma.cc/CEY8-34QS (explaining that because "[h]ealth apps, wearables, social media platforms, and many other technology companies" are often not covered by HIPAA, these companies "can most often legally share, license, and sell users' health data . . . without users' knowledge or consent").

72 Justin Sherman, *Your Health Data Might Be for Sale*, SLATE (June 22, 2022), https://perma.cc/Z4QW-AT8L.

harassment.[73] In contrast, those who collect data directly from consumers often have a contractual relationship in the form of, for instance, terms of use. For example, Google recently settled (for an undisclosed amount) a consumer privacy lawsuit "claiming [that Google] secretly tracked the internet use of millions of people who thought they were doing their browsing privately," violating federal wiretapping and California privacy laws.[74] The determinative issue in this dispute was whether "Google had made a legally binding promise not to collect users' data when they browsed in private mode."[75] Unlike most data brokers, first-party data collectors often have contractual relationships that limit the ways in which they collect data. In contrast, while data brokers can still be punished for data breaches[76] and violations of state law,[77] they are otherwise largely safe to transact as they please.

Fourth, and finally, it should temper some privacy concerns if consumers meaningfully consented to their data being exploited in these ways. But troves of evidence show that consumers do not meaningfully consent to such collection and often have no clue that their data is being used in these ways.[78] Furthermore, even if express consent were extractable, a broad swath of literature has also found that "[c]onsumer consent is not an effective, administrable, or viable approach to the regulation of commercial surveillance."[79] The power to define what constitutes consumer consent lays in the hands of companies, which often—like "many laws and bills around the country"—define it "as a person simply using an application or service that has a privacy policy."[80] To use

---

73 Sherman, *Sensitive Data*, *supra* note 7, at 3.

74 Jonathan Stempel, *Google Settles $5 Billion Consumer Privacy Lawsuit*, REUTERS (Dec. 29, 2023), https://perma.cc/J4UH-WHVN.

75 *Id.*

76 One example is the credit agency Equifax, which paid $575 million after a data breach that exposed the personal and financial information of nearly 150 million people after failing to fix a critical vulnerability in their database. *See Equifax to Pay $575 Million as Part of Settlement with FTC, CFPB, and States Related to 2017 Data Breach*, FED. TRADE COMM'N (July 22, 2019), https://perma.cc/2PBA-KECQ.

77 *See, e.g.*, Justin Sherman, *Examining State Bills on Data Brokers*, LAWFARE (May 31, 2022) [hereinafter Sherman, *Examining State Bills on Data Brokers*], https://perma.cc/PY8Q-H5RL. *See generally* Solon Barocas & Helen Nissenbaum, *Big Data's End Run Around Anonymity and Consent*, *in* PRIVACY, BIG DATA, AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 44 (Julia Lane, Victoria Stodden, Stefan Bender & Helen Nissenbaum eds., 2014) (arguing that procedural protections, such as required disclosure, are not enough to protect consumer privacy).

78 *See, e.g.*, OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE 59–78 (2014).

79 *See, e.g.*, Sherman et al., *supra* note 37, at 6.

80 *Id.*

these technologies, consumers are virtually forced to accept terms of use that quietly permit the collection and sale of their data. A regime supported by consumer consent alone, therefore, remains inadequate for the data brokerage context.

II. CURRENT REGULATIONS FAIL TO PROPERLY ACCOUNT FOR DATA'S UNIQUE PROPERTIES

Given the dismal status quo and dearth of federal regulation, courts might illuminate the way forward. But turning to courts introduces new hurdles for litigants. To establish standing in the traditional civil suit, litigants must satisfy, among other requirements, the showing of a "concrete and particularized" injury.[81] But establishing that brokering harms are concrete can prove difficult for at least two reasons. First, privacy harms are often intangible. While "intangible injuries can nevertheless be concrete," the Supreme Court has acknowledged that "tangible injuries are [ ] easier to recognize."[82] Second, some of data's unique properties as a quasi-public good incentivize its repeat exploitation, which makes it difficult to ascertain precisely how much exploitation has occurred. Before embracing a turn to the courts, it is important to acknowledge the inherent properties of data that underlie this conceptual difficulty. Appreciating these properties will aid courts in calculating the legally relevant harms from brokering. Relatedly, these properties that encourage data's repeated exploitation also illustrate why monitoring the data brokerage industry should not be left to the free market.

This Part proceeds as follows. Section A explores some of the relevant properties of data that encourage its repeated exploitation—data is nonrivalrous, practically nonexcludable, and synergistic. These properties often make it difficult for courts to accurately calculate magnitudes of harm. Section B introduces and explores de-anonymization, a strategy that capitalizes on data's synergistic nature to link otherwise anonymized data back to the person from whom it was collected. De-anonymization presents another variable in the harm calculation that courts have not yet meaningfully considered.

---

[81] *See* Lujan v. Defs. of Wildlife, 504 U.S. 555, 560 (1992).

[82] Spokeo, Inc. v. Robins, 578 U.S. 330, 340 (2016).

A. Data's Properties as a Quasi-Public Good Encourages Its Repeated Exploitation

Data is a unique quasi-public good, in that it is nonrivalrous, practically nonexcludable, and synergistic. These characteristics enable and encourage firms, including data brokers, to exploit it repeatedly.

1. Data is nonrivalrous.

First, data is nonrivalrous—one person's use does not necessarily deprive another person of simultaneous or subsequent use. It "can be used by infinitely many people without depriving the original owner of the use of their property."[83] With the click of a button, an original owner can copy their data and distribute that copy to another party. Doing so does not deprive them of their original copy. In theory, then, the same data can belong to an infinite number of owners at once. Furthermore, the harm that each copy inflicts would depend on who owns that copy and what they did (or plan to do) with it. Some uses do not harms at all and promote competition and consumer welfare by, for instance, empowering its users to "increase efficiency and innovate."[84] In sum, data's nonrivalrous nature presents a conceptual challenge: How can courts calculate the harm caused by an infinitely duplicatable good, assuming that the magnitude of harm not only increases, but increases in varying amounts with each additional copy?

The challenge is not insurmountable in practice, though it is not trivial either. One might think that natural market forces of competition provide the proper incentives for parties not to wantonly copy and distribute their data. And this is true to some extent. Google and Amazon each own treasure troves of consumer data. But there seems to be very little incentive for them to share that data with one another, lest this gives the other party a leg up in the market. Exclusivity is valuable. So while it is true that infinitely many parties can theoretically own the same copy of data at once, in practice the value of data depends at least somewhat on its exclusivity.

---

[83] *Intellectual Property*, CORNELL L. SCH. LEGAL INFO. INST., https://perma.cc/F6LN-WSDV.

[84] YAN CARRIÈRE-SWALLOW & VIKRAM HAKSAR, INT'L MONETARY FUND, THE ECONOMICS AND IMPLICATIONS OF DATA: AN INTEGRATED PERSPECTIVE 5 (2019).

Private parties' competitive incentives to gatekeep data are not sufficient to solve the problem, however. While exclusivity offers one reason for a party not to distribute data, there are just as many (if not more) compelling reasons to distribute it.[85] An economically rational party has little reason to keep data to itself, unless utilizing the data privately confers a comparative advantage. But there are many situations where there is no such comparative advantage. For instance, the same data can be used for different, nonoverlapping purposes. DoorDash might want its consumer data to improve its share of control over the food delivery market. But a market research firm might want that same data for an entirely different purpose—to assess the impact of promotions on consumer behavior. While the fruits of exclusivity might convince DoorDash not to sell its consumers' data to Grubhub, that same aversion to selling would apply less strongly to entities seeking to use the same data for different purposes. The existence and success of the data broker market tends to prove this observation—data brokers collect data not for personal use, but to resell.

In sum, data is nonrivalrous—it is infinitely duplicable. A data broker can sell data to one party, then turn around and sell that same data to another party. The nonrivalrous nature of data creates incentives for data brokers to fully exploit the data they own, then transfer it to other parties for them to exploit anew. More crucially, these entities internalize the benefits of selling data but outsource the negative externalities of doing so to the consumers whose privacy is being harmed. From a consumer's perspective, entities sell their data again and again, resulting in compounding damage to consumer privacy each time. Courts may face difficulty in calculating those harms with adequate precision.

---

[85] Data's nonrivalrous nature distinguishes its market's properties from those of tangible goods. Ignoring the negative externalities that sales of data create, the "welfare-optimal solution" for the data market might be "to price the data at zero so that it could be used as much as possible to maximize its potential value." STUCKE, *supra* note 53, at 153.

For personal data, this suggests no privacy at all. That cannot be right. Still, a solution cannot go so far as preventing *all* sales of personal data. A balancing act arises between privacy and the benefits of a functioning data market—such as competition. On one hand, companies "need access to personal data" to "compete and innovate." *Id.* at 165–66. On the other, the repeated mining of data imposes negative externalities on privacy. *See id.* A good solution needs to balance those competing tensions.

2. Data is practically nonexcludable.

Second, in practice, data is often nonexcludable. "[E]ase of misappropriation . . . distinguish[es] [data] from many other forms of property."[86] Control over informational access is more difficult to regulate than access to tangible goods. Comprehensively tracking the ownership and transfer of data is practically impossible. And some uses of data, such as using data to train machine-learning models, are impossible to reliably identify.[87] Unless a data broker advertises their services, regulators have insufficient methods to trace a broker's sales. State disclosure laws might remedy these information asymmetries somewhat, but existing laws are not so granular as to require disclosure on a transaction-by-transaction basis.[88] Because personal data is collected and sold broadly, such granular disclosure mandates would likely be overly burdensome for both regulators and the market. For instance, Xandr, Microsoft's advertising and analytics subsidiary, discloses that "it may send data to 1,647 other companies."[89] The challenge, then, is determining how courts can make do with imperfect knowledge. The nonexcludable nature of data combined with the sheer size of the data ecosystem renders tracking the provenance of data difficult. This inspires little confidence that courts can do better.

3. Data is synergistic.

Third, data—particularly personal data—is synergistic to a much greater degree than most tangible goods. Data synergy describes how "data from multiple sources . . . , when combined, is more valuable than any of the sources were on their own."[90] This synergy makes the impact of selling data difficult to predict and

---

86 U.S. DEP'T OF JUST. & FED. TRADE COMM'N, ANTITRUST GUIDELINES FOR THE LICENSING OF INTELLECTUAL PROPERTY § 2.1 (2017).

87 *See, e.g.*, Andersen v. Stability AI Ltd., 2023 WL 7132064, at *8 (N.D. Cal. Oct. 30, 2023) ("The [ ] problem for plaintiffs is that it is simply not plausible that [all training data] used to train [the model] was copyrighted . . . or that all [of the model's outputs] rely upon (theoretically) copyrighted [training data], and therefore *all* [o]utput images are derivative images." (emphasis in original)).

88 *See, e.g.*, CAL. CIV. CODE § 1798.99.80 (West 2024).

89 Sabine Zimmer, Ron Bradley & Tom Garrubba, *Real-Time Bidding: Technology or Data Breach?*, SHARED ASSESSMENTS (May 20, 2022), https://perma.cc/4V2Y-SAR2.

90 Sarah Higginson, Marina Topuzi, Carlos Andrade-Cabrera, Ciara O'Dwyer, Sarah Darby & Donal Finn, *Achieving Data Synergy: The Socio-Technical Process of Handling Data*, *in* ADVANCING ENERGY POLICY 63, 64 n.3 (Chris Foulds & Rosie Robison eds., 2018).

perceive. To be sure, consolidating tangible property also frequently produces synergistic outcomes. In fact, property law is often designed to move goods to their highest and best user, and presumably, the highest and best user often owns assets that become more valuable when a synergistic good is obtained. The concept that the whole can be greater than the sum of its parts seems intuitive for tangible goods.

The same concept applies to data. In fact, three features of data make it a uniquely synergetic good, perhaps even more so than tangible goods. First, data is tabular—datasets are typically organized in tables of rows and columns—which makes it easy to consolidate datasets. Two distinct brokers can separately collect data from the same subject. If those databases are combined, data belonging to the same subject can be aggregated with the click of a button. Second, machine-learning algorithms allow owners of data to efficiently extract inferences between a dataset's different attributes. And third, the aggregation of personal data creates more apparent externalities. What underpins this final observation is the fact that personal data exists not in a vacuum; rather, it belongs to a data subject. Sales and aggregation of datasets implicate the privacy of the subjects represented in those datasets.

To illustrate, consider a buyer who owns the following database representing some personal data of imaginary subjects:

TABLE 1

First Name	**Birth Year**	**Occupation**
Alex	2001	Film Critic
Billy	2000	Park Ranger
Claire	1999	Scientist

The buyer, seeking to expand their dataset, purchases a similar database that collects different attributes[91]:

TABLE 2

First Name	**Last Name**	**Favorite Food**
Darcy	Dixon	Steak
Billy	Butler	Pizza
Claire	Clark	Pie

Having both in hand, the buyer could easily run an algorithm to "join" the two datasets to produce the following consolidation[92]:

TABLE 3

First Name	**Last Name**	**Birth Year**	**Occupation**	**Favorite Food**
Alex		2001	Film Critic	
Billy	Butler	2000	Park Ranger	Pizza
Claire	Clark	1999	Scientist	Pie
Darcy	Dixon			Steak

The tabular nature of data enables owners to mine its synergies efficiently and fully.

Furthermore, not only is data synergistic, but it can be synergistic in unpredictable ways. Notice how the resulting dataset exposes more information about Billy and Claire compared to Alex and Darcy. Billy and Claire appeared in both original datasets, so it makes sense that the consolidated data would reflect more knowledge about them.[93] But buyers and sellers do not always know exactly who is represented in the datasets they buy. And even if the buyers and sellers do know, an outside observer (including a potential regulator) does not. This creates an information asymmetry. The sometimes unpredictable synergy of data

---

91 First name, last name, and favorite food are this dataset's "attributes." Darcy, Billy, and Claire are its "data subjects."

92 For purposes of simplifying the illustration, the assumption here is that it is the same Billy and Claire represented in both datasets.

93 Even if there is no overlap in the subjects represented in two databases, owning more raw data can help train algorithms that are better at making inferences to predict missing data. *See* Sherman et al., *supra* note 37, at 4 (explaining how data brokers can use data to derive additional data about individuals).

exacerbates the uncertainty of calculating brokering harms.[94] This example illustrates two premises. First, data becomes more valuable in the aggregate. Second, the privacy costs of aggregation depend on the data already (or soon to be) in the buyer's possession.

If it is aggregation that results in the parade of horribles, perhaps those horribles might be avoided by somehow limiting aggregation. Observe that merging the datasets required entries from both datasets to be linked to a common data subject. The subjects' first names were the "identifier" that allowed such a linkage. If regulators want to strike a balance between permitting sales of data but avoiding unpredictable outcomes that arise from data aggregations, why not just remove the "first name" column before selling the dataset, thus neutering the buyer's ability to merge the datasets? This concept of removing or obscuring identifiers is known as "anonymization." The law has recognized this mechanism and has carved out safe harbors for firms who anonymize their data before selling it.[95] But as explained in the next Section, the synergistic nature of data ironically also enables sophisticated de-anonymization techniques, where firms can reidentify individuals by cross-referencing two datasets, even when such identification is not achievable when looking at the two datasets separately.

B. Anonymization Is an Imperfect Solution

A data broker might anonymize their data before offering it on the free market,[96] but not for altruistic reasons. Rather, they would likely do so to procure the "get-out-of-jail-free card" that "nearly every information privacy law or regulation grants . . . to those who anonymize their data."[97] To begin, the data broker

---

94 *See* Fluitt et al., *Data Protection's Composition Problem*, *supra* note 30, at 291 ("[I]t has become exceedingly difficult (and in many cases impossible) to predict how fast privacy degrades with each new data use.").

95 *See* Ohm, *Broken Promises*, *supra* note 29, at 1704.

96 For an overview of anonymization techniques, see Ohm, *Broken Promises*, *supra* note 29, at 1711–16, and see generally Boris Lubarsky, Note, *Re-Identification of "Anonymized" Data*, 1 GEO. L. TECH. REV. 202 (2017).

97 Ohm, *Broken Promises*, supra note 29, at 1704. Federal privacy statutes offer safe harbors for those who anonymize. *See, e.g.*, 45 C.F.R. § 164.514(a) (2023) ("Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information."). California's CCPA expressly does not cover

might delete identifiers. First and last names are a good start. Phone numbers and email addresses are next. After tediously removing attributes that seem to identify its data subject, the broker offers the dataset for sale. But are the identities of these subjects truly protected? Or can a motivated actor expose their identities through seemingly innocuous attributes such as their favorite movies?

A motivated actor would likely be able to de-anonymize the data. "Reidentification has become horrifyingly easy."[98] In 2006, AOL published a collection of twenty million web searches from 650,000 people. Despite replacing their names with random numbers, reporters "very quickly linked the searches to specific people."[99] Two years later, researchers "famously matched 500,000 Netflix users' 'anonymized' movie ratings against IMDb" and identified not only the users' identities but sensitive information about them like their political preferences.[100] And "[w]hen researchers examined a data set from the New York City government, again without names, of every single taxi ride in the city," they were able to identify over 91% of the taxis and could even "classify drivers' incomes."[101] In all three instances, motivated actors employed de-anonymization techniques to unmask data subjects' identities.

So how does de-anonymization work? "De-anonymization" refers not to a particular technique, but to a class of techniques that use external information to make inferences about an anonymized dataset.

"deidentified" information. CAL. CIV. CODE § 1798.145(a)(1)(F) (West 2024). The CCPA defines deidentified information as information that "cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer" provided that a business that uses deidentified information takes four operational and organizational steps to ensure that such information is not reidentified or disseminated. *Id.* § 1798.140(ab). Even overseas, Europe's comprehensive data privacy law excludes "anonymized" data from its reach. *See* Council Regulation 2016/679, 2016 O.J. (L 119) 5.

98 Justin Sherman, *Big Data May Not Know Your Name. But It Knows Everything Else.*, WIRED (Dec. 19, 2021), https://www.wired.com/story/big-data-may-not-know-your-name-but-it-knows-everything-else/ [hereinafter Sherman, *Big Data Knows Everything Else*].

99 *Id.*

100 *Id.*

101 *Id.*

To demonstrate how it might work, consider the following dataset:

TABLE 4

Name	**Age**	**Location**
Jack	12	Up the hill
Jill	7	Up the hill

Meanwhile, the seller owns this dataset:

TABLE 5

Name	**Height**	**Action**
Jack	5'1"	To fetch a pail of water
Jill	4'6"	To fetch a pail of water

If the seller sells their dataset to the buyer and does not anonymize the data before selling, the consolidated dataset would look like this:

TABLE 6

Name	**Height**	**Age**	**Location**	**Action**
Jack	5'1"	12	Up the hill	To fetch a pail of water
Jill	4'6"	7	Up the hill	To fetch a pail of water

However, if the seller operates in a jurisdiction offering safe harbors for anonymizing data, they might wish to anonymize their data before they sell it. To do so, the seller might obscure the dataset like so:

TABLE 7

Name	**Height**	**Action**
?	5'?	To fetch something
?	4'?	To fetch something

Now, the buyer of that dataset is unable to merge the two datasets. What would have formerly served as a common attribute between the datasets—their names—is now obscured. Jack and Jill are safe.[102]

Not so fast. A clever buyer could try to uncover the identities of people represented in the dataset even though some data is missing. A motivated actor could seek out external information to try to de-anonymize this data. They could, for instance, ask around to see if anyone "fetched something" recently to learn that two siblings went up the hill to fetch a pail of water. They might then find a family photo on Facebook and estimate that Jack is roughly five feet tall and that Jill is a little shorter, thus increasing the chance that the data belongs to the siblings. And with that, the attacker has linked Jack and Jill's identities back to the "anonymized" dataset and can merge the datasets as they please.

In the grand scheme of things, this illustration presents a crude and relatively unsophisticated de-anonymization technique. But sophisticated actors can employ algorithms to sort through massive troves of data and automate the de-anonymization process. An algorithmic technique might take attributes from the dataset that, if taken alone, would not identify Jack and Jill, and combine them. It is wholly possible that the now-combined attribute serves as a valid identifier. A core principle animates that possibility: "a small number of data points about an individual, none of which are uniquely identifying, are collectively equivalent to an identifier."[103] Even if a seller removes Jack and Jill's names, that would do little in hiding their identities if (1) they were the only two people that went "up the hill" to "fetch something," and (2) a motivated actor somehow acquires that knowledge.

Studies suggest that "over 99% of Americans could be correctly re-identified from any dataset using 15 demographic attributes."[104] Simply knowing someone's birthdate, zip code, and gender is enough to identify roughly 87% of all people in the United States.[105] "While there might be a lot of people who are in their

---

[102] *But see Jack and Jill*, WORDS FOR LIFE, https://perma.cc/S6KN-PGEX ("Jack fell down and broke his crown, [a]nd Jill came tumbling after.").

[103] Arvind Narayanan & Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets: A Decade Later 1 (2019) (unpublished research paper) (available at https://perma.cc/AV4W-6TAX).

[104] Nick Wells & Leslie Picker, *'Anonymous' Data Might Not Be So Anonymous, Study Shows*, CNBC (July 23, 2019), https://perma.cc/RQN6-WSAW.

[105] Brian Hayes, *Uniquely Me!*, 102 AM. SCIENTIST 106, 106 (2014).

thirties, male and living in New York City, far fewer of them were also born on January 5, are driving a red sports car and live with two kids . . . and one dog."[106] Every additional piece of data whittles down the possible matches.

Over a decade ago, privacy scholar Professor Paul Ohm, in an illuminating article, emphasized how and why these anonymization techniques fell flat, and how the law has not meaningfully responded.[107] It is not clear that this has changed significantly since then. Waiting does no favors for privacy, as computing capabilities and the raw amount of data only increase with time. As late as 2019, computer scientists showed that approximately 99.98% of anonymized data may be capable of re-identification and, as explored above, the risks of re-identification are heightened when data is aggregated.[108]

So why not just use better anonymization techniques? Many of the sensationalist stories of de-anonymization are the result of unsophisticated or poorly executed anonymization techniques. For example, in the New York taxis story, researchers were able to "backtrack from [ ] badly generated hash codes."[109] To simplify, the dataset relied on a secret code to translate names into gibberish. But the secret code was too weak, and the researchers were able to break it and uncover the original data. Still, that represented more of a technical oversight than a fundamental methodological problem—a stronger secret code would have fared better. So why isn't the solution simply a better secret code?

That solution, while attractive, faces three challenges. First, the effectiveness of anonymization depends not just on how well broker *A* anonymizes its data. It also—and arguably even more so—depends on how well *others* anonymize their data. Put differently, anonymization across the data trade is only as strong as its weakest link. De-anonymization has a pseudocommutative property.[110] To demonstrate, assume that there are two data brokers in the data trade. Assume that broker *A* fails to anonymize, but

---

106 Wells & Picker, *supra* note 104.

107 *See* Ohm, *Broken Promises*, *supra* note 29, at 1743 (arguing that Congress has focused legislation on anonymization, but anonymization "should no longer be considered to provide meaningful guarantees of privacy").

108 *See generally* Luc Rocher, Julien M. Hendrickx & Yves-Alexandre de Montjoye, *Estimating the Success of Re-identifications in Incomplete Datasets Using Generative Models*, 10 NATURE COMMC'N 3069 (2019); Narayanan & Shmatikov, *supra* note 103.

109 Sherman, *Big Data Knows Everything Else*, *supra* note 98.

110 An operation is commutative if changing the order of the operands does not change the result. For example, addition is commutative: $1 + 2 = 2 + 1$.

broker *B* anonymizes well. It makes no difference whether broker *A* fails to anonymize or if broker *B* fails to anonymize, as long as (1) at least one of them fails to do so adequately, and (2) a motivated actor has access to both datasets. Recall that de-anonymization techniques often cross-reference external data with the data they are trying to de-anonymize. Hence, an attacker can use broker *A*'s dataset as the "external data" and cross-reference broker *B*'s dataset with it. The same is true vice versa.[111]

Second, the raw amount of data accumulates over time, while de-anonymization techniques also increase in complexity. A dataset that today may be deemed properly anonymized might not be in the future. Perhaps data collected in the future enables an attacker to cross-reference it with older data and infer a person's identity. Or a new de-anonymization technique can reverse the efficacy of older anonymization techniques (before new techniques are developed in response). These risks can be difficult to predict, and the consequences are difficult to detect.

Third, and relatedly, observe that de-anonymization is a collective risk. The risk that my identity is linked with data does not depend solely on the data that belongs to me. It also depends on whether data can be linked to others. When others' identities are exposed, the process of elimination brings an attacker ever so slightly closer to finding my identity. For instance, say that an attacker has narrowed this dataset as belonging to one of two potential people—John and Jane:

TABLE 8

Name	**Zodiac Sign**	**Favorite Color**
?	Pisces	Blue

[111] This is a classic collective action problem. Industry standards that establish best practices for anonymization can partially remedy the problem. *See, e.g.*, Luk Arbuckle, *A New Standard For Anonymization*, IAPP (Mar. 14, 2023), https://perma.cc/KJ8W-E362 (explaining a new privacy standard that works to identify and mitigate various risks across the lifecycle of deidentified data). But industry standards without the threat of corresponding legal sanctions cannot eliminate every weak link. So responsible data brokers, even if they comply with industry best practices, are forced to bear the risks that sloppy data brokers produce. Considering that, it seems unreasonable to require that data brokers anonymize their data to the point where it is immune from de-anonymization, because doing so would likely also completely deplete the value of the data. Responsible data brokers bear the brunt of the costs that irresponsible data brokers generate.

Then, the attacker acquires this dataset:

TABLE 9

Name	**Zodiac Sign**
John	Aquarius

Because John is an Aquarius, he cannot be the mystery person represented in the first dataset. The attacker can now infer that Jane is a Pisces whose favorite color is blue. The broader principle is that data collected belonging to person *A* often renders the data of person *B* more easily identifiable—and at scale, this facilitates de-anonymization.

As this Part has shown, it is difficult to accurately conceive of the harms that arise from the sale of data. Data's inherent properties enable and incentivize entities to repeatedly sell and aggregate it. Each sale creates externalities of varying magnitudes. De-anonymization further complicates the harm inquiry. The efficacy of anonymization is undercut by irresponsible sellers who anonymize sloppily or fail to anonymize at all. The harms that arise are volatile and difficult to predict. How are courts supposed to conceptualize these harms, let alone calculate them within an acceptable degree of error?

As the next Part explores, current regulatory frameworks have failed to account for these difficulties. Instead, they largely analyze data brokers in insolation from one another. In doing so, they fail to appreciate a vital observation: the harm inflicted by the sale of data depends on what other data exists in the market. As the next Part explores, that oversight obscures and systematically undermeasures the harms that sales of data inflict.

III. CONCEPTUALIZING PRIVACY HARMS: *FTC V. KOCHAVA*

Despite the inherent limitations that data's informational form imposes on regulating data brokers, some novel enforcement actions seem promising. This Part explores a budding area of privacy litigation—the FTC's suits against data brokers pursuant to its § 5 authority to prevent "unfair or deceptive acts or practices."[112] After providing an overview of § 5 and its scope, this Part explores an ongoing lawsuit between the FTC and Kochava, a

[112] 15 U.S.C. § 45(a)(1).

data broker.[113] In August 2022, the FTC sued Kochava in an Idaho district court for selling location data that enabled the tracking of people at reproductive health clinics, places of worship, and other sensitive locations.[114] Walking through *Kochava*'s procedural history illustrates (1) how courts conceive of data brokering harms and (2) that courts have not meaningfully responded to the risks of de-anonymization in calculating harm—a risk exacerbated by data's nonrivalrous and nonexcludable nature.

A. Regulation of Data Brokers and the "Unfair Acts and Practices" Standard Under the FTCA

Section 5 authorizes the FTC to prohibit "unfair . . . acts or practices in or affecting commerce."[115] A company can run afoul of § 5 if, for instance, it misleads consumers by failing to comply with statements in its posted privacy policies or makes material changes to privacy policies without providing adequate notice to consumers.[116] As discussed earlier, the FTC has applied this

---

[113] *See generally* FTC v. Kochava Inc., 671 F. Supp. 3d 1161 (D. Idaho 2023).

[114] *FTC Sues Kochava*, *supra* note 56.

[115] 15 U.S.C. § 45(a)(1). Section 5 of the FTCA is better known for a similar prohibition of "[u]nfair methods of *competition* in or affecting commerce," *id.* (emphasis added), which serves as the statutory basis for the FTC's antitrust enforcement. The FTC recently released a policy statement announcing an intention to wield this power more expansively. *See generally Policy Statement Regarding the Scope of Unfair Methods of Competition Under Section 5 of the Federal Trade Commission Act*, FED. TRADE COMM'N (Nov. 10, 2022), https://perma.cc/65YD-HJRC.

Discussing the policy statement, FTC Chair Lina Khan explained how "[i]n passing [the FTCA], Congress [ ] tasked the FTC with identifying the range of methods of competition that qualify as unfair, . . . recogniz[ing] they could not specify them all prospectively." *See* Lina M. Khan, *Section 5 in Action: Reinvigorating the FTC Act and the Rule of Law*, 11 J. ANTITRUST ENF'T 145, 149 (2023). While some have argued that "Section 5 should be read merely as extending Sherman Act enforcement authority to the FTC," Khan disagreed, arguing instead that the "straightforward reading of the statute" is that Section 5 furnishes the FTC with the power to "challenge a host of unlawful business practices not covered by the other antitrust laws." *Id.*

It is not made expressly clear to what extent, if any, the policy statement's analysis of what constitutes an "unfair method of competition" also extends to "unfair acts or practices in or affecting commerce." In fact, Khan expressly noted that she "use[s] Section 5 as shorthand for the unfair methods of competition prohibition" and "do[es] not address unfair or deceptive acts or practices." *Id.* It may be that the FTC is trying not to bite off more than it can chew. But even so, the logic undergirding the FTC's expansive move here supports a similarly expansive reading of the FTC's ability to enforce against unfair acts and practices.

[116] *See* 15 U.S.C. § 45(n); *see also Privacy and Security Enforcement*, FED. TRADE COMM'N, https://perma.cc/MU8B-9R4G.

power to reach everything from telemarketing schemes[117] to deceptive shaving cream commercials.[118]

Recent events strongly suggest that § 5 reaches at least some data brokering activities. The FTC is currently testing these theories in federal court and its administrative tribunals. In a January 2024 administrative proceeding, the FTC reached its first settlement with a data broker—X-Mode—concerning the collection and sale of sensitive location data.[119] Mere weeks later, the FTC settled another agency proceeding against InMarket Media under a similar theory.[120] It is that same theory that underlies the ongoing civil suit against Kochava. This Comment suggests that these theories have underexplored the risks of de-anonymization. Properly exploring these risks would strengthen the FTC's theories of liability. Furthermore, understandably, the FTC is currently picking off low-hanging fruit: data brokers who totally fail to anonymize. Importing a more nuanced understanding of de-anonymization may extend the FTCA's reach beyond data brokers who totally fail to anonymize to data brokers who anonymize but nonetheless do so ineffectively when considering other data available in the ecosystem.

The term "unfair" in § 5 is intentionally broad.[121] In the process of enacting the FTCA, Congress "explicitly considered[ ] and rejected[ ] . . . reduc[ing] [ ] ambiguity . . . by enumerating the particular practices to which [§ 5] was intended to apply."[122] "Instead, Congress authorized the FTC to use its expertise in guiding the law's application and development in different contexts."[123] This state of affairs held for the first eighty years of the Act—"Congress remained mostly on the sidelines and let the FTC develop the meaning of unfairness through policy statements and agency adjudications."[124] "But in 1994, spurred by growing criticisms of the FTC's liberal use of Section 5, Congress amended the

---

117 *See* FTC v. Windward Mktg., Inc., 1997 WL 33642380, at *1 (N.D. Ga. Sept. 30, 1997).

118 *See* FTC v. Colgate-Palmolive Co., 380 U.S. 374, 374 (1965).

119 *See Order Prohibits Data Broker X-Mode Social and Outlogic from Selling Sensitive Location Data*, FED. TRADE COMM'N (Jan. 9, 2024), https://perma.cc/TE4E-5VCT.

120 *See FTC Finalizes Order with InMarket Prohibiting It from Selling or Sharing Precise Location Data*, FED. TRADE COMM'N (May 1, 2024), https://perma.cc/4CQ2-EXXR.

121 *See Kochava*, 671 F. Supp. 3d at 1169 ("If those terms seem broad, they are intentionally so.").

122 FTC v. Sperry & Hutchinson Co., 405 U.S. 233, 239–40 (1972) (citing S. REP. NO. 63-597, at 13 (1914)).

123 *Kochava*, 671 F. Supp. 3d at 1169.

124 *Id.*

FTCA and added [a three-pronged test] to limit the FTC's authority to deem acts and practices 'unfair' under Section 5[ ]."[125] This three-pronged test still stands today. An act or practice is unfair where it (1) causes or is likely to cause substantial injury to consumers, (2) cannot be reasonably avoided by consumers, and (3) is not outweighed by countervailing benefits to consumers or to competition.[126] Each of these prongs calls for elaboration.

First, an unfair act or practice must cause or be likely to cause substantial injury to consumers. While substantial injury usually involves monetary or physical harms,[127] courts have recognized that § 5's use of "injury" is "not limited to tangible injuries" and includes "intangible invasion[s] of a legally protected interest."[128] But a limiting principle exists: "[t]rivial or merely speculative harms are typically insufficient for a finding of substantial injury."[129] For instance, scholars have recognized that, "in all but the most extreme cases, individual dignitary harms are likely not considered injuries that the FTC may address under its unfairness authority."[130] Privacy harms sit somewhere in between but are closer to a cognizable substantial injury than not. Invasion of privacy is generally considered a concrete harm, not merely a form of "mental distress" harm.[131] Applying those principles, courts have recognized that the disclosure of sensitive

---

125 *Id.* (citing FTCA Amendments of 1994, Pub. L. No. 103-312, § 9, 108 Stat. 1691, 1695).

126 *See Federal Trade Commission Act Section 5: Unfair or Deceptive Acts or Practices*, *in* CONSUMER COMPLIANCE HANDBOOK, 1, 8 [hereinafter *Section 5 Guidance*] (chapter available at https://perma.cc/RZT6-W5D3).

127 *Id.*

128 *Kochava*, 671 F. Supp. 3d at 1173.

129 *Section 5 Guidance*, *supra* note 126, at 8 ("Emotional impact and other more subjective types of harm will not ordinarily make a practice unfair.").

130 Andrew D. Selbst & Solon Barocas, *Unfair Artificial Intelligence: How FTC Intervention Can Overcome the Limitations of Discrimination Law*, 171 U. PA. L. REV. 1023, 1042 (2023); *see also* Mark MacCarthy, *New Directions in Privacy: Disclosure, Unfairness and Externalities*, 6 I/S: J.L. & POL'Y FOR INFO. SOC'Y 425, 484 (2011) ("Emotional distress, mental anguish, loss of dignity and other harms are not [categorically] ruled out . . . , but they must be effects that all or most or reasonable persons would construe as genuine harms.").

131 *See, e.g.*, RESTATEMENT (SECOND) OF TORTS § 652B (AM. L. INST. 1977) ("One who intentionally intrudes, physically or otherwise, upon the solitude or seclusion of another or his private affairs or concerns, is subject to liability to the other for invasion of his privacy, if the intrusion would be highly offensive to a reasonable person."); *see also, e.g.*, Patel v. Facebook, Inc., 932 F.3d 1264, 1273 (9th Cir. 2019) (recognizing that the collection of "otherwise unknowable" information "implicate[s] privacy concerns" (quotation marks omitted) (first quoting Carpenter v. United States, 138 S. Ct. 2206, 2215 (2018); then quoting Riley v. California, 573 U.S. 373, 393 (2014))).

medical information is an actual concrete harm "even without economic or other tangible harm."[132]

Furthermore, the phrase "likely to cause substantial injury" has been read to "incorporate[ ] both the probability and the magnitude of harm, so that a lower probability will suffice if the magnitude of the harm is sufficiently great."[133] "An act or practice that causes a small amount of harm to a large number of people may be deemed to cause substantial injury. An injury may be substantial if it raises a significant risk of concrete harm."[134] This flexible balancing between probability and harm is particularly useful when applied to data-related harms. Upon, say, a data breach, the probability that an attacker targets any particular data subject is relatively small. But those risks, when aggregated, are far more likely to support a finding of substantial harm.

Second, "[c]onsumers must not reasonably be able to avoid the injury."[135] Here, courts focus on whether the act or practice "interfere[d] with [the consumer's] ability to effectively make [informed] decisions."[136] Intentional deception offers the most concrete example: hiding the price of a product or service until after the consumer has committed to purchasing it would prevent the consumer from making an informed purchasing decision. Courts also consider whether the act or practice is unduly coercive. "[A]gencies will not second-guess the wisdom of [ ] consumer decisions" on a case-by-case basis, but will ask the general question of whether behavior "creates or takes advantage of an obstacle to the free exercise of consumer decision making."[137]

Third, "[t]he injury must not be outweighed by countervailing benefits to consumers or to competition."[138] Pairing this prong with the substantial-injury prong resembles a sort of cost-benefit balancing. Courts do not seem to apply strict limits on what can be considered a "countervailing benefit," and the FTC's guidance documents say as much: "the injury must not be outweighed by

---

[132] *In re* Horizon Healthcare Servs. Inc. Data Breach Litig., 846 F.3d 625, 635 (3d Cir. 2017).

[133] Brief of the Federal Trade Commission at *17, LabMD, Inc. v. FTC, 2017 WL 562771 (11th Cir. Feb. 9, 2017) (No. 16-16270).

[134] *Section 5 Guidance*, *supra* note 126, at 8.

[135] *Id.*

[136] *Id.*

[137] *Id.*

[138] *Id.*

any offsetting consumer or competitive benefits that are also produced by the act or practice."[139] For instance, some benefits might include lower prices or a more diverse, wider availability of goods and services. Courts also consider the costs incurred when determining the remedies for allegedly unfair acts.

This Comment will focus primarily on the first prong—substantial injury—because that prong has proved the most difficult for the FTC in the context of regulating data brokers. The case of *FTC v. Kochava*,[140] to which the next Section turns, illustrates this difficulty.

B. *Kochava* Illustrates the Substantial-Injury Requirement

In August 2022, the FTC sued Kochava Inc., a data broker, in an Idaho district court for selling location data that enabled the tracking of people at reproductive health clinics, places of worship, and other sensitive locations.[141] The FTC sought to enjoin sales of this location data under the theory that such sales were "unfair,"[142] because such data could "expos[e] [individuals] to threats of stigma, stalking, discrimination, job loss, and even physical violence."[143]

Kochava moved to dismiss the complaint under Federal Rule of Civil Procedure 12(b)(6), and in May 2023, the district court granted Kochava's motion without prejudice to the government.[144] The court ruled that while the FTC adequately pled two prongs of the unfairness test—that consumers could not reasonably avoid the alleged harms and that these harms were not outweighed by countervailing benefits—it failed to adequately plead the first prong, a likelihood of substantial consumer injury.[145]

The FTC had put forth two theories of consumer injury: (1) a direct theory of harm—that "the disclosure of consumers' sensitive location information *itself* constitutes substantial injury to consumers' right to privacy"[146]—and (2) a secondary theory of harm—that Kochava's location-data sales "could enable third parties to track consumers' past movements to and from sensitive

---

139 *Section 5 Guidance*, *supra* note 126, at 8 (emphasis added).
140 671 F. Supp. 3d 1161 (D. Idaho 2023).
141 *FTC Sues Kochava*, *supra* note 56.
142 *See id.*
143 *Id.*
144 *Kochava*, 671 F. Supp. 3d at 1168.
145 *See id.* at 1171–76.
146 *Id.* at 1171 (emphasis added).

locations and, based on inferences arising from that information, inflict secondary harms including 'stigma, discrimination, physical violence, [and] emotional distress."[147] While the court acknowledged that both theories could theoretically constitute a substantial injury, it held that the FTC had not plausibly pled that the alleged injury rose to the requisite level of substantiality.[148]

Regarding the direct harm theory, the court held that the alleged privacy intrusion was not "sufficiently severe to constitute" substantial injury.[149] The court found that "three factors lessen[ed] the severity of the alleged privacy injury."[150] First, Kochava sells data that is not *facially* sensitive. Rather, sensitive data "can be ascertained only by inference[s]," which are often unreliable.[151] Second, the data Kochava sells can be "accessible through other, lawful means," such as by observing a person's movement in public.[152] Third, the Commission failed to "generally indicate[ ] how many device users may suffer privacy intrusions," which is important "because the substantiality of a consumer injury depends, in part, on the number of consumers injured."[153] While the court discussed these three attenuating circumstances in the context of evaluating the Commission's direct-harm theory, the same circumstances also seem to weaken its secondary-harm theory. Particularly, the court noted how inferring sensitive data was often unreliable. If sensitive data is accessible only via unreliable inferences, it follows that a third party would have a more difficult time targeting any particular individual.

Regarding the secondary-harm theory, the FTC failed to allege that "consumers [were] suffering or [were] likely to suffer such secondary harms" and had "only allege[d] that secondary harms [were] *theoretically* possible."[154] That was not enough. The

---

147 *Id.*

148 *See id.* at 1171–77.

149 *Id.*

150 *Kochava*, 671 F. Supp. 3d at 1175.

151 *Id.*; *see also id.*:

> [G]eolocation data showing that a device visited an oncology clinic twice in one week could reveal that the device user suffers from cancer. Or it may instead reveal that the person has a friend or family member who suffers from cancer. Or that the person is a pharmacist or is in the business of selling or maintaining medical devices. The point is that the FTC does not actually claim that Kochava is disclosing private information, but rather that it is selling data from which private information might be inferred.

152 *Id.*

153 *Id.*

154 *Kochava*, 671 F. Supp. 3d at 1171 (emphasis added).

court explained that the FTC must "go one step further and allege that Kochava's practices create a 'significant risk' that third parties will identify and harm consumers."[155]

In November 2023, the FTC filed an amended complaint featuring enhanced pleadings that Kochava causes a likelihood of substantial injury.[156] The new complaint included additional details about Kochava's alleged unfair practices, asserting that Kochava's data can be used to trace consumers' movements to locations that are "sensitive and personal."[157] It further emphasized that the data Kochava sells is not anonymous.[158] And it highlighted real-world instances where individuals were targeted—the Catholic priest who was outed, and a data broker who sent targeted advertisements about abortion to the broker's "abortion-minded women."[159] Kochava promptly moved to dismiss the amended complaint under Federal Rule of Civil Procedure 12(b)(6), arguing that the FTC had not "cured the deficiencies" that had led to its first dismissal.[160]

In February 2024, the court denied Kochava's motion to dismiss.[161] The court found that both of the FTC's harm theories were adequate. For the direct harm theory, the court emphasized that, while "inferences based on geolocation data, alone, *can be* unreliable," the FTC's new allegations allege that "Kochava itself makes inferences about consumers, rather than simply providing raw data from which its customers could make inferences."[162] For the secondary harm theory, the court noted how, unlike the original complaint, the amended complaint "contain[ed] allegations that the targeting of consumers based on geolocation data 'has and does occur.'"[163]

This is, of course, a win for proponents of privacy. Alongside recent events, specifically the FTC's settlements with other data

---

155 *Id.* at 1172 (quoting FTC v. Neovi, Inc., 604 F.3d 1150, 1157 (9th Cir. 2010)).

156 *See generally* Amended Complaint, FTC v. Kochava, 2024 WL 449363 (D. Idaho Feb. 3, 2024) (No. 2:22-CV-00377).

157 *In Amended Complaint, FTC Alleges Kochava, a Data Broker, Is Collecting, Using and Disclosing "Massive Amounts" of Precise Geolocation Data*, PROSKAUER (Nov. 6, 2023), https://perma.cc/C46D-LEQW.

158 Amended Complaint, *supra* note 156, at *28.

159 *Id.* at *28–29.

160 *See* Memorandum in Support of Motion to Dismiss Amended Complaint, *Kochava*, 2024 WL 449363 (No. 2:22-CV-00377).

161 *See generally* Order Denying Motion to Dismiss First Amended Complaint, *Kochava*, 2024 WL 449363.

162 *Id.* at *5 (emphasis in original).

163 *Id.* at *9.

brokers, the current posture of the *Kochava* litigation affords some confidence that § 5 unfairness theories have bite as applied to data brokers.

But at the same time, victory should not be declared too soon. The court's rejection of Kochava's motion to dismiss raises additional questions. For instance, the court found that it was meaningful that Kochava itself made inferences instead of offering raw data. It is not entirely clear why. Is it because, in doing so, Kochava had done much of the heavy lifting for bad actors to commit these secondary harms? Is it, instead or additionally, because Kochava's inference making dirtied their hands, thus conferring some sense of heightened culpability for any secondary harms that ensue? The answer is important for data brokers seeking legal compliance. More crucially, the answer will determine just how far these § 5 theories can go against data brokers. If the FTC's pleadings are taken as truth, Kochava, in the grand scheme of things, is low-hanging fruit. Their anonymization efforts are lackluster, to say the least, and they openly advertise their inference making.[164] Would a data broker that anonymizes, but does so poorly, be subject to liability? The discussion in Part II suggests that the answer should at least sometimes be yes. Recall that the same data can be magnitudes more useful (and hence more harmful to data subjects) to an owner if they own the right data to aggregate it with. Thus, depending on the entire network of available data, imposing liability there might be normatively desirable.

Kochava illustrates the difficulty—for both courts and litigants—of quantifying harms that arise from data brokers' operations. But more importantly, it leaves open a big question: How should courts assess the risk that a motivated actor successfully makes harmful inferences? *Kochava* sketches out some lower bounds: the data broker might run afoul of § 5 if it makes the inappropriate inferences itself and offers them on the market. But what about data brokers who offer raw data? Or those who anonymize with varying degrees of success? As Part II articulated and the next Part expounds, looking at the entire network of available

---

[164] Amended Complaint, *supra* note 156, at *22–23 ("Kochava . . . boast[s] that the Kochava Collective contains 'other points to connect to and securely solve for identity.'"); *id.* at *23 (alleging that Kochava advertises that customers can use its database to identify the consumer's name, address, phone number, email address, gender, age, yearly income, economic stability, marital status, education level, and more).

data will help courts with this line drawing problem. Part IV offers a framework that incorporates de-anonymization and other network risks in judicial conceptions of harm.

IV. A PATH FORWARD: INCORPORATING NETWORK HARMS IN THE PRIVACY CALCULUS

Kochava illustrates how courts conceive of data brokering harms. Under the conventional approach, brokering harms are treated as discrete harms severable from other activity occurring in the larger network.[165] A court could theoretically calculate the harm that the sale of a dataset inflicts by looking only at *that* dataset. How much privacy does *that* dataset intrude? What secondary harms does *that* dataset enable? At no point would a court need to look beyond the four corners of the dataset to answer those questions.

But that conception of privacy harm overlooks a crucial intuition: the magnitude of the harm that arises from data brokering depends on what other data is accessible in the network. De-anonymization, for instance, often relies on cross-referencing *external* data to make internal inferences. Furthermore, a motivated actor can purchase multiple datasets from different brokers, employ de-anonymization techniques to overcome barriers to aggregation, and then aggregate as they please. The aggregated dataset would implicate a more substantial privacy interest and present a higher risk of secondary harms than if any one of those datasets were to be owned in isolation.

The current methodology thus systematically underestimates the magnitude of brokering harms. To capture network harms in the harm calculus, courts should look beyond simply identifying the direct or secondary harms that arise from and only from the specific broker's sale. Instead, courts must analyze the projected consequences of that sale if combined with other available data on the market. Incorporating network harms in this way results in both a more descriptively accurate and normatively desirable regime. Such incorporation is not without objections, however, and this Part addresses some in turn before ultimately concluding that courts should still incorporate network

---

[165] *See, e.g.*, *id.* at *22 ("Kochava's data is not anonymized and is linked or easily linkable to individual consumers. Indeed, Kochava actively markets its ability to link consumers' real names, addresses, email addresses, and phone numbers to sensitive information, including their gender, marital status, and age.").

harms. Next, this Part explores a nonexhaustive list of factors that a court can consider when measuring network harms. Finally, this Part applies the proposed framework to *Kochava*.

A. Courts Should Recognize Network Harms

Judicial recognition of these "network harms" serves two objectives, one descriptive and one normative. First, doing so provides a more accurate descriptive account of the harms that arise from data brokering. The traditional approach operates on the assumption that the harm that one broker inflicts can be compartmentalized from those inflicted by another broker.[166] But that is not always the case. Recall that the risk that a given dataset is de-anonymized depends substantially on the existence and accessibility of relevant data external to that dataset.[167] How harmful broker *A*'s sales are depends on whether broker *B* is selling compatible data which, if aggregated, would enable de-anonymization. In other words, it is relevant whether a motivated actor can discern from external information exactly who "went up the hill" to "fetch a pail of water." Relatedly, looking only to an individual sale overlooks the possibility that a buyer purchases and aggregates multiple datasets, an act which heightens the risk and magnitude of harm even if no de-anonymization occurs.[168] Recognizing network harms harmonizes the judicial conception with the axiom that the whole is greater than the sum of its parts.

The second objective is normative. Incorporating network harms would extend the reach of litigants' theories of liability to data brokers whose activities are not grossly negligent on their face but are nevertheless harmful. One could imagine a more sympathetic data broker than Kochava whose dataset offerings are partially anonymized but nevertheless serve as the key to exposing someone's identity. In those instances, it may be normatively desirable to enjoin sales, even if there are other brokers practicing equally middling anonymization standards that simply do not result in the same harms by the pure happenstance of what data is available on the market. A network-harms approach is consequentialist and more concerned with preventing harm than it is vindicating innocent mens rea.

---

166 Professor Justin Sherman drew a similar distinction between laws that target the underlying ecosystem of data brokerage and those that do not. *See* Sherman, *Examining State Bills on Data Brokers*, *supra* note 77.

167 *See, e.g.*, *supra* Part II.B.

168 *See supra* Part II.B.

This Section now turns to three objections to incorporating network harms. The first objection contends that network harms do not occur or occur so infrequently as to constitute no more than a rounding error. No one but tenure-seeking computer science professors are de-anonymizing, the argument goes.[169] Furthermore, even if sensitive information is being exposed via data brokers, these harms are largely attributable to *individual* data brokers. Therefore, introducing network harms would obscure and unnecessarily complicate the inquiry.

There are three responses to this objection. First, this Comment has assumed thus far that a hypothetical aggregator is a third party. But recall that there is significant cross-pollination in the data brokerage industry—"[s]everal . . . data brokers share the same sources," and the majority of the studied brokers "buy from or sell information to each other."[170] There are thus strong financial incentives for these brokers to shop for compatible data and aggregate at scale. Second, most of the nightmarish incidents making headlines, such as that involving the former Catholic priest, are surgical attacks on specific individuals carried out for a particular purpose. It is not preposterous to assume that those same attackers would be motivated enough to shop around and combine data to achieve their objectives. To be sure, if an attacker could obtain all the data they needed from a single source, that would certainly save them the trouble of collecting data from multiple sources and aggregating it. But data brokers seem to be receiving increasing public scrutiny—states are slowly passing data broker laws,[171] and the FTC is hammering down on the worst offenders. Eventually, then, the supply of these one-stop shops for sensitive data is likely to dry. Network harms would certainly be relevant then. Finally, recall that "it is startlingly easy to reidentify people in anonymized data."[172] The risk of de-anonymization increases over time as data accumulates, algorithms improve, and computing capabilities become stronger and more accessible to the public.

The second objection contends that incorporating network harms complicates an already attenuated inquiry of measuring

---

[169] Ohm has referred to this objection as the "Myth of the Superuser." *See* Ohm, *Broken Promises*, *supra* note 29, at 1730.

[170] *See* FTC, DATA BROKERS, *supra* note 48, at 14.

[171] *See supra* note 24.

[172] Ohm, *Broken Promises, supra* note 29, at 1730 ("[M]ost people who have taken a course in database management or worked in IT can probably replicate this research using a fast computer and widely available software like Microsoft Excel.").

intangible privacy harms and predicting the risk that tangible secondary harms arise. Even if network harms exist, attempting to account for them could inflate the margin of error beyond workable limits. This objection can also be overcome. Section 5's substantial-injury prong is a *risk* inquiry that is necessarily speculative and predictive. Some uncertainty is thus inevitable. Incorporating network harms may exacerbate this uncertainty, but it is certainly not introducing it in the first instance. Nor does it exacerbate uncertainty beyond workable limits. This objection applies with significantly less force to litigants like the FTC who represent a large number of people. In the garden-variety civil suit with one plaintiff and one defendant (e.g., a privacy tort suit), a plaintiff seeking to incorporate network harms would need to tie those harms to their specific facts and establish causation. Mass adoption of network-harms analysis is thus unlikely in the private litigation context where the acceptable margin of error is smaller. It is probably not viable to make general assertions of network and de-anonymization harms if there is no way to anchor them to the plaintiff or defendant. In § 5 litigation, by contrast, the margin of error is bigger because the relevant inquiry is harm done to the entire American public. Network harms would thus likely fit into that wider acceptable margin of error.

The final objection contends that incorporating network harms inequitably punishes brokers who responsibly anonymize. Recall that de-anonymization is commutative[173]: if dataset *A* cross-referenced with dataset *B* results in de-anonymization, it follows that the same is true when dataset *B* is cross-referenced with dataset *A*. Even a beautifully anonymized dataset can be de-anonymized if paired with an irresponsibly anonymized dataset. Under a strict conception of network harms, the brokers selling each dataset are equally liable—their datasets, when consolidated, ultimately result in aggregation. This objection contends that it is inequitable to hold both the responsible and irresponsible brokers equally culpable.

There are two responses. First, due to resource constraints, the FTC and other litigants will likely pursue the worst offenders, which partially insulates responsible brokers. Second, in a functional data ecosystem, irresponsible anonymizers simply inflict more harm than their responsible counterparts in the aggregate, even though de-anonymization is commutative. To be sure, if the

---

[173] *See supra* note 110 and accompanying text.

brokerage network consisted of two brokers selling datasets that could be aggregated with each other's, their sales are equally harmful from a network-harms perspective. But the brokerage network is more elaborate than that. Imagine, for instance, a brokerage network with ten total brokers, where nine are responsible anonymizers and the remaining broker is an irresponsible anonymizer. In mixing and matching their datasets, the nine responsibly anonymized datasets can only be aggregated with the single irresponsibly anonymized dataset. In stark contrast, the irresponsible broker's dataset can be aggregated with all nine other datasets on the market. A network-harms approach thus imposes liability proportionate with the harms inflicted by each broker's sale.

This discussion suggests that, on balance, incorporating network harms into the calculus is both workable and normatively desirable. The next Section turns to the proposed framework.

B. The Framework

Before presenting the framework in earnest, it is worth discussing whether a balancing framework is truly more workable than a series of categorical rules. Courts could model the latter approach on statutory analogs enumerating categories of data that are simply off limits to sell or buy. HIPAA, for instance, restricts the sale of sensitive medical information under certain circumstances.[174] But there are two reasons to prefer a balancing framework over categorical rules. First, de-anonymization risks depend on context and thus lie on a spectrum, which pushes in favor of a flexible standard. Second, bright line rules tend to be a poor fit here because the harms inflicted by secondary uses of data depend on *how* that data is actually used.[175] A comprehensive, categorical solution risks being overinclusive—the transfer of sensitive data is not per se harmful. Medical data, for instance, could be immensely useful to medical research but equally useful to an insurance company seeking to discriminate against patients based on their medical history. And it also risks being underinclusive—since anonymization can be ineffective on the margins, brokers are not properly incentivized to transact with due care. If

---

[174] *See* 45 C.F.R. § 160.103 (2023) (defining "covered entities" and "business associate[s]" under HIPAA's purview).

[175] *See* Daniel J. Solove, *Data Is What Data Does: Regulating Based on Harm and Risk Instead of Sensitive Data*, 118 NW. U. L. REV. 1081, 1084 (2024) ("To be effective, privacy law must focus on harm and risk rather than on the nature of personal data.").

we wish to maximize capturing the social benefits of data brokerage, a categorical approach requires exceptions. But exceptions for what purposes? Or for whom? Those exceptions might threaten to swallow the rule. On balance, these considerations make a more flexible and pragmatic standard, rather than a categorical rule, more attractive.

This Comment's proposed framework serves as a tool for courts to quantify the harms that data brokers inflict when they sell a dataset. Its factors, while nonexhaustive, serve as a starting point to aid courts in determining when the sale of a dataset produces an inappropriate risk of harm. Unlike the conventional model, this framework incorporates network harms, particularly the threat of de-anonymization. This framework recognizes two main categorizes of harm: (1) inherent harms and (2) network harms. The first category encompasses the harms that exist inherently in selling the data, regardless of what other data brokers in the network are selling. For example, the sale of poorly anonymized, sensitive medical information is damaging on its own, regardless of activities occurring in the market as a whole. Put differently, inherent harms reflect those that hypothetical attackers could inflict wielding only that dataset—and nothing more. The inherent harms that arise from the sale of a dataset increase with four factors:

1. *Anonymization*. How well the dataset is anonymized serves as a baseline proxy for how much harm a dataset inflicts when sold.

2. *Sensitivity*. Some data, even personal data, is inherently more sensitive than other types of data. This judgment is partially subjective, but objective lines can be drawn. For example, location data seems objectively more sensitive than data on someone's eye color.

3. *Dataset size*. The larger the dataset is, the more data points there are to mine inferences and the more weak points there are for an attacker to de-anonymize.

4. *Contemporaneousness*. The contemporaneousness of the dataset also matters. Old, stale data is likely to be less harmful.

The second category encompasses harms that arise or are exacerbated by synergistic data sold by other brokers in the network. For example, take the same poorly anonymized, sensitive medical information. If another broker sells data that enables an

attacker to de-anonymize that information, that constitutes a network-level harm. These network harms increase according to:

1. *Consumer overlap*. Many de-anonymization techniques rely on cross-referencing external data on a consumer to figure out which consumers are represented in obscured datasets. The extent to which dataset *B* can be cross-referenced to de-anonymize dataset *A* requires that both datasets have data on the same consumer. Broadly, courts should consider where the data was sourced from. It is unlikely that plaintiffs can find this information in granular form. But they can make solid estimates. The likelihood of consumer overlap increases, for example, when the datasets were collected from the same geographic area or when the datasets were purchased from the same first-party collector (such as an app).

2. *Attribute overlap*. The risk of de-anonymization harm increases when two datasets sell data that covers the same attributes. This is because the more that the attributes overlap, the more likely an attacker can pinpoint entries between the two datasets to compare. Concerns are heightened when the attribute serves as an identifier (i.e., each attribute identifies a single person) like an email address or a phone number.

3. *Temporal overlap*. The risk of de-anonymization harms increases if the two datasets share data that was collected at the same time. This is especially true for data points that tend to change over time. For example, someone's race does not change over time. Conversely, someone's medical history is very likely to change.

Courts can weigh these factors to estimate the risks of de-anonymization. Certainly, satisfying all three factors is not required for a court to conclude that de-anonymization risks are unduly high. For example, if the court finds significant consumer and attribute overlap but no temporal overlap whatsoever (e.g., imagine that dataset *A* is ten years older than dataset *B*), it might still conclude that de-anonymization risks are too high. The court's conclusion would be particularly strong if the data in question was static (e.g., race) as opposed to dynamic (e.g., favorite food). The next Section applies the framework to *Kochava*.

C. Applying the Framework to *Kochava*

To maximize the reach of § 5 by incorporating network harms in the calculus, the FTC's secondary harm theories—that selling

data enables potential harms, such as stalking, by third parties[176]—seem to be the ideal vehicle. Alas, evaluating the risks of de-anonymization constructively assumes that there is a motivated secondary actor attempting to de-anonymize data. But showing to courts' satisfaction that these harms are substantial and likely to occur will be more difficult. Recall that the FTC must still allege that "consumers are suffering or are likely to suffer such secondary harms."[177] Alleging that "secondary harms are theoretically possible" is not enough.[178] To be sure, in *Kochava*, the FTC adequately pled its secondary harm theory.[179] But as discussed in Part III, Kochava was relatively low-hanging fruit—it made inappropriate inferences itself and virtually no meaningful anonymization efforts. The bar will be much harder to clear if the FTC wishes to target data brokers whose activities are less flagrant but nonetheless dangerous.

Moreover, even in its amended complaint, the FTC did not identify any actual secondary harms that arose directly from Kochava's sales.[180] Any identified secondary harms were instead caused by *other* data brokers and were included in the complaint only to show that such harms were possible. The FTC instead argued that harm was likely to occur because (1) Kochava's location data exposed visits to sensitive locations (which in turn exposed individuals to secondary harms), (2) the dataset lacked any controls on who could access this sensitive data, and (3) the dataset lacked sufficient guarantors of anonymity, such that users could link data back to its individuals.[181] Perhaps that is the entire point of § 5—a one-to-one causal link need not be drawn directly between an entity's activities and a specific instance of harm. Indeed, in *Kochava* itself, the FTC adequately pled a § 5 case in its amended complaint without citing to concrete harms attributable to Kochava itself. But that raises the question: Will courts buy those same arguments against a more sympathetic broker, such as one that did not blatantly make inappropriate inferences or totally fail to anonymize? In those circumstances, will the FTC's

---

176 *See Kochava*, 671 F. Supp. 3d at 1171.

177 *Id.*

178 *Id.*

179 *See* Order Denying Motion to Dismiss First Amended Complaint, *supra* note 161, at *9 ("Unlike the original Complaint, the Amended Complaint contains allegations that the targeting of consumers based on geolocation data 'has and does occur.'").

180 *Id.* ("Kochava responds [to the complaint] that none of the FTC's 'anecdotes' involve its own data.").

181 Amended Complaint, *supra* note 156, at *28–29.

current strategy of showing "a likelihood of substantial harm" by pointing to general de-anonymization disasters caused by other data brokers be enough?

An approach that considers network and de-anonymization risks on a case-by-case basis will temper these worries. The proposed framework would allow the courts to incorporate de-anonymization into the court's conception of harm. Applying the factors for inherent harm:

1. *Anonymization*. Kochava does not make any efforts to anonymize the data it sells.[182]

2. *Sensitivity*. Kochava sells location data, which is highly sensitive. Additionally, its data also identifies ethnicity, gender identity, date of birth, minor status, number of children, political association, and marital status.[183]

3. *Dataset size*. Kochava's datasets are large. It claims to have data on "over 300M unique individuals in the US" with up to "300 data points that can be tied to those profiles."[184]

4. *Contemporaneousness*. Kochava's data is recent and frequently updated.[185]

Next, applying the factors for network harms, we find that it is very likely that an attacker could cross-reference Kochava's data to de-anonymize other datasets:

1. *Consumer overlap*. Kochava collects data from at least "10,000 apps globally,"[186] which makes it extremely likely that other data brokers sell consumer data that overlaps with Kochava's.

2. *Attribute overlap*. Kochava collects a significant number of different attributes, and many serve as identifiers (e.g., name, address, email address, phone number).[187] The raw number of different attributes Kochava collects makes it likely that these attributes also overlap with the attributes in other data brokers' datasets.

---

[182] *Id.* at *13.

[183] *Id.*

[184] *Id.* at *12 (quotation marks omitted).

[185] *See id.* at *8:

> Kochava has asserted that it offers "rich geo data spanning billions of devices globally." It has further claimed that its location data feed "delivers raw latitude/longitude data with volumes around 94[ ] [billion] geo transactions per month, 125 million monthly active users, and 35 million daily active users, on average observing more than 90 daily transactions per device."

[186] Amended Complaint, *supra* note 156, at *16.

[187] *Id.* at *12.

3. *Temporal overlap*. Because Kochava collects data from 125 million monthly active users and has been operating since 2011,[188] it is very likely that the data Kochava collects overlaps temporally with the data of other brokers.

Observe that the framework highlights the risk that Kochava's poorly anonymized data can be used to de-anonymize other datasets. In the FTC's amended complaint, the Commission emphasized that Kochava had made no efforts to anonymize its data, and that such data "c[ould] be and is used to identify consumers and sensitive information about them."[189] In addition to these allegations, the Commission could strengthen its complaint by referencing other data brokers that provide datasets likely to include individuals overlapping with Kochava's data subjects. For example, if Kochava were to collect the location data about people living in a specific geographic area, the Commission could point to another data broker that sells purchase-history data from the same geographic area. Doing so would recognize that data can be consolidated, and so the harm that arises if Kochava does not anonymize its dataset extends beyond the information present in its own dataset. Incorporating network harms into the calculus would capture those unpredictable side effects of transacting in a seamlessly interconnected data ecosystem.

CONCLUSION

Congress has failed to properly respond to the data broker epidemic. A turn to the next best thing—the courts—has followed. This Comment suggests that litigants and courts, by treating data brokering as producing discrete harms, may be underestimating the actual harm that occurs. The magnitude of the harm that arises from one broker's activities depends on what other data is available in the larger network. This Comment recommends a framework for courts to conceive of these network harms and promotes its implementation. The FTC's § 5 enforcement actions against data brokers offer a practical means to implement this framework.

---

[188] Jeff Richardson, *The Next Generation Mobile Measurement Partner (MMP)*, KOCHAVA (Mar. 18, 2022), https://perma.cc/6SS2-FR6D.

[189] Amended Complaint, *supra* note 156, at *4.

The law is said to always be at least five years behind emerging technologies.[190] Here, where there is a risk of substantial privacy harm, we do not have the luxury of waiting. A blow to a large data broker like Kochava could invigorate industry pushback that may even culminate in a federal data broker law. Until then, leveraging the FTC's § 5 authority to recognize network harms provides a proactive and effective path forward.

[190] Manav Tanneeru, *Can the Law Keep Up With Technology?*, CNN (Nov. 17, 2009), https://perma.cc/QMP4-L94D.